ARMY RESEARCH LABORATORY

U.S. Army Research Laboratory (ARL) Corporate Dari Document Transcription and Translation Guidelines

by Luis Hernández and Sherri Condon

ARL-TN-0512 **October 2012**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when no longer needed. Do not return to the originator.

Army Research Laboratory
Adelphi, MD 20783-1197

ARL-TN-0512 **October 2012**

U.S. Army Research Laboratory (ARL) Corporate Dari Document Transcription and Translation Guidelines

Luis Hernández and Sherri Condon
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
October 2012	Final	

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
U.S. Army Research Laboratory (ARL) Corporate Dari Document Transcription and Translation Guidelines	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Luis Hernández and Sherri Condon	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
U.S. Army Research Laboratory ATTN: RDRL-CII-T 2800 Powder Mill Road Adelphi, MD 20783-1197	ARL-TN-0512
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report provides guidelines for the transcription and translation of text content found in document page images obtained from printed materials. It is intended to create a conditional facsimile of the textual areas of interest found in a document page and its associated English translation in standard text file format.

15. SUBJECT TERMS

Transcription, Translation, guidelines, ground truth, Optical character recognition, OCR, Machine Translation, MT, Dari

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Luis Hernández
Unclassified	Unclassified	Unclassified	UU	22	19b. TELEPHONE NUMBER (*Include area code*) (301) 394-4301

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

1. Introduction

Developers of human language technologies in a foreign language often require printed document page images, their corresponding machine-readable text, and its associated translation from a foreign language into a target language in order to train, test, and evaluate optical character recognition (OCR) and machine translation (MT) embedded systems (*1–4*). When producing data to test OCR capabilities, it is necessary to consider the OCR system typical workflow—preprocess, filter, segment, analyze, and recognize—and the issues associated with OCR generation processes. Production of full ground truth for a document image is a laborious process that requires many decisions concerning the ways that page segments will be represented and the granularity of the representation. Even if the goal is to produce only a text reference, as is here, decisions must be made about how the original text will be represented. Similarly, representation of text for operations like MT requires the system to correctly sequence the words from one segment to the next. For example, a sentence or paragraph may begin at the bottom of one column, but end at the top of an adjacent column. These kinds of information must be captured in a representation conducive to translation systems usage from a representation obtained from the text in the image. Obtaining data to support the research and development of these technologies into systems is expensive and requires a systematic approach to capture and document such important resources' content variety and linguistic features.

2. Purpose

This report provides guidelines for the transcription and translation of text content found in document page images obtained from printed materials. The source of the printed material content comes from scanned page images, page photographs, or portable document format (pdf) files. The applicability of this report is limited to Dari text content. The intent of this report is to create a conditional facsimile of the textual areas of interest found in a document page and its associated English translation in standard text file format. The transcription guidelines herein only cover text regions of interest in images and are not intended to create full layout document page image ground truth data.

3. Document Image Transcription

The transcription guidelines call for preserving line breaks from the original document page image in order to facilitate using transcriptions to create full ground truth for the documents at a

future time. Document images that contain only a table of contents or only a few lines of text (such as a dedication or a title) should not be transcribed. Transcribers should keep a record of pages that are not transcribed for this reason. Also, document images that contain text that is illegible due to degradation of the image should not be transcribed and should also be recorded in the document that lists pages not transcribed.

3.1 Transcription Format

The electronic files from the transcription task should be plain text files in UTF-8 encoding. The transcribed text size and font family should be 14 pt Times Roman. The file should end at the final character in the document image with *no* additional carriage return: the hidden end-of-file marker should immediately follow the final character in the document image. Transcribers should not begin a new empty line at the end of the file.

3.2 File Naming Conventions

Each resulting transcription filename should correspond to a document page image filename, except for the addition of the suffix *"_trs*.txt" at the end of the filename. For example, if the document image file is named "file.pdf," the transcription file should be named "file_trs.txt."

3.3 File Internal Content Structure Conventions

Each line in the transcription file should correspond to a text line in the document image file. To preserve the text line, each line should end with a carriage-return/linefeed marker, Lines in the file should have the following format where [tab] represents a 2-space tab delimiter and square brackets represent positions in the line (not included in the transcription):

[image filename][tab][page number xxxx][tab]Line 0001[tab]first line of file
[image filename][tab][page number xxxx][tab]Line 0002[tab]second line of file
[image filename][tab][page number xxxx][tab]Line 0003[tab]third line of file
[image filename][tab][page number xxxx][tab]Line 0004[tab]fourth line of file

and so on until all the text lines present in the image are transcribed.

> **Example 1:** A single page image file named "*Document1"* contains a total of 15 lines of text. The resulting transcription filename would be "Document1_trs.txt" displaying the file structured content as follows when opened in a text editor:
>
> Document1 0001_1 0001 text line content
> Document1 0001_1 0002 text line content
> . . . .
> . . . .
> . . . .
> Document1 0001_1 0015 text line content

The image filename should match the filename of the document image only. The above example illustrates the case for document page images containing one page per file. In the case of image

files containing two facing document printed pages, the page number that is recorded in the content structure should reflect this fact. Likewise, pdf files may contain multiple pages in a single file and should also reflect this association. The following three cases may occur:

1. If the document page image file captures a single page of the text only, use "0001_1", as shown in Example 1, to associate the text content that is transcribed. The page number used might not be the first page of the printed document: the page number only reflects the sequence of the images whose text content is transcribed.

2. If the document image file contains two facing pages of a book, the first file from the document should be numbered, for example, "*0001_1*" for the first facing page and "*0001_2*" for the second facing page. Subsequent image files from the same document should be numbered sequentially while preserving this schema, i.e., "*0002_1, 0002_2, 0003_1, 0003_2*", and so on.

   Example 2: The document image file contains two facing pages of a book and each page contains 3 lines:

   ```
   Document1  0001_1  0001     text line content
   Document1  0001_1  0002     text line content
   Document1  0001_1  0003     text line content
   Document1  0001_2  0001     text line content
   Document1  0001_2  0002     text line content
   Document1  0001_2  0003     text line content
   ```

3. If the document image file is a pdf file containing multiple single pages from the document, the page number should correspond to the page number in the pdf file as it is displayed when the file is opened in Adobe Acrobat Reader. The pdf page numbers begin with 1 on the first page of the document and continue sequentially until the end of the file. The page numbers used in the content structure should reflect that fact. For example, if the transcription begins on the first page of the pdf document, then the transcription lines would appear as in Example 1. (Documents consisting of multiple single-page images will be identifiable from sequence numbers in the image filenames.) If the transcription begins on the 4th page of the document, then the lines will appear as in Example 3.

   Example 3: The document image file contains multiple pages, each corresponding to a page in Adobe Acrobat Reader. The transcription begins on the 4th page according to Adobe Reader. "n" corresponds to the nth page and "x" represents the last text line found at the end of each page in the pdf file:

   ```
   Document1  0004_1  0001     text line content
   Document1  0004_1  0002     text line content
         .       .        .          .
         .       .        .          .
         .       .        .          .
   Document1  0004_1  000x     text line content
   ```

```
Document1   0005_1 0001   text line content ⟵ Next page in Adobe Reader
Document1   0005_1 0002   text line content
    .       .       .         .
    .       .       .         .
    .       .       .         .
Document1   0005_1 000x   text line content
    .       .       .         .
    .       .       .         .   ⟵ Intervening pages
    .       .       .         .
Document1   000n_1 0001   text line content ⟵ Final page transcribed
Document1   000n_1 0002   text line content
    .       .       .         .
    .       .       .         .
    .       .       .         .
Document1   000n_1 000x   text line content
```

3.4 Representing the Text Lines in the Page

Each line in the transcription file should correspond to a line in the document image file. The characters in every line of the original document should be transcribed, except as specified below, including page numbers, headers, and titles. Text captions that label or title graphic elements should be transcribed, and all numbers should be transcribed. Numbers that appear in Indic form should be transcribed in the same form.

Non-text elements, such as separator lines, bullet symbols, or graphical elements such as photographs, illustrations, and other artwork, should not be transcribed or replicated. An exception is text presented as calligraphy: if the text is horizontal and readable, then it should be transcribed (and translated). But if the text in calligraphy is vertical or forms a pattern such as a circle, then it should be treated as a graphic element and should not be transcribed. Elements that are not part of the primary text such as handwritten annotations or stamps should not be transcribed.

The lines of the transcription should reflect the order of the text on the page beginning at the top-right of the page and continuing to the bottom of the page. Blank lines should not be used to reflect vertical spacing of the lines in the original document. Transcription files should not contain blank lines.

Each line should consist of the words and characters that occur in the corresponding line in the original document in the order that they appear. Centering, indentation, and other horizontal spacing of words should not be represented in the transcription file. (Spacing in tables will be represented with the tab character as explained below.) Each transcription line should begin at the rightmost position in the line, unless the line begins with a symbol or punctuation character, which is any keyboard character that is not a letter or a number. In this case, a space should be inserted at the beginning of the line.

For pages consisting of more than one column of text, the lines in the transcription should correspond to the lines in the columns. The lines should be ordered beginning with the lines in

the rightmost column with the last line of that column followed by the first line of the next column to the left and continuing until all columns have been transcribed (see example 4 in figure 1). No symbols should be used to indicate the beginnings or ends of columns.

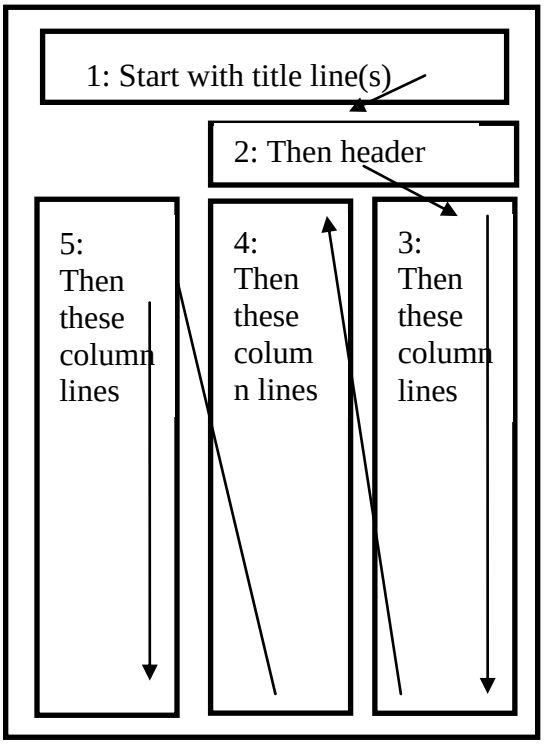
For instances in which text wraps around graphic elements, the order of the lines should reflect the reading order. If text wraps around a box of text, text that occurs above or to the right of the box should be transcribed before text in the box and all the text lines in the box should be transcribed before any text lines that are left or below the box (see example 5 in figure 1).

Transcription of tables is an exception to the general ordering rules. Table transcription should preserve each row's columns using 2-space tabs. A carriage-return/linefeed should be inserted at the end of each table row. Example 6 (in figure 1) demonstrates the order of transcription lines for tables. The order illustrated in Example 6 should be used regardless of the presence of border lines in the table. The same conventions should also be used for any material in side-by-side columns when there is a relationship between the items on the same line, such as a list of names and phone numbers.

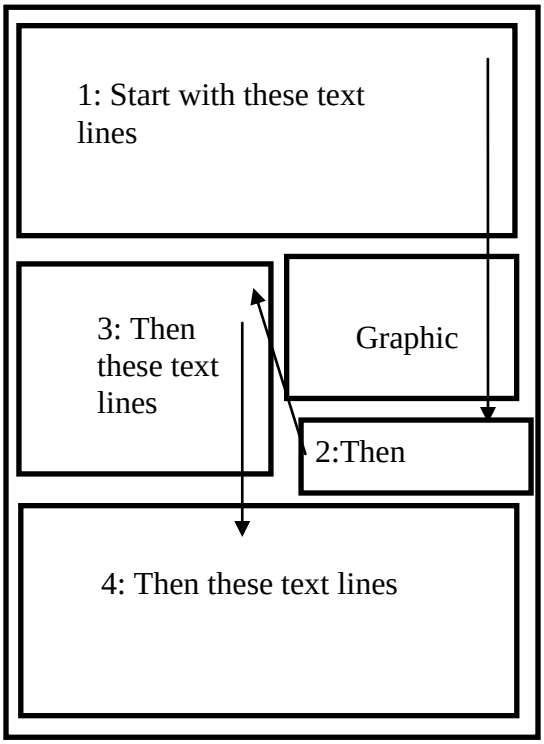
For complex formats, lines should be ordered to conform as much as possible to the following three principles, which are listed here in order of importance:

1. Preserve the reading order of the text content.
2. Text that is higher on the page is transcribed before text that is lower on the page. (For this ordering convention, superscripts and subscripts are treated as if they were on the same line as the words they modify, e.g., in a sequence like *…word*^{2} the "^{2}" is transcribed as "word^{2}" even though it is technically above "word" in the page.)
3. Text that begins farther right on the page is transcribed before text that occurs farther left.

Example 4: Columns



Example 5: Text Wrapped around Graphic



Example 6: Table Image

Dari Table Title on Top

Dari Table Column Label 3	Dari Table Column Label 2	Dari Table Column Label 1
Dari table cell 3	Dari table cell 2	Dari table cell 1
Dari table cell 6	Dari table cell 5	Dari table cell 4

Table Caption Below

Transcription

Dari Table Title on Top
Dari Table Column Label 3[**tab**]Dari Table Column Label 2[**tab**]Dari Table Column Label 1
Dari table cell 3[**tab**]Dari table cell2[**tab**]Dari table cell 1
Dari table cell 6[**tab**]Dari table cell 5[**tab**]Dari table cell 4
Dari Table Caption Below

Figure 1. Examples for transcribing columns, wrapped text around an image, and tables.

3.5 Representing Characters in the Text

Characters should be reproduced as they appear in the text, except as specified below. Some characters are used in handwriting or typewritten documents that cannot be reproduced using standard keyboard mappings. For example, when documents were produced on typewriters, it was possible for the typist to mistakenly use the wrong form of a character such as the initial form of *heh* instead of the medial form (see Example 7). Modern software does not permit this kind of error. In these cases, the character should be consistently substituted with an appropriate character that is distinct from other characters in use in the corpus, and this substitution should be documented in a documentation file that accompanies the corpus.

Example 7: original: **فهرست**

automated rendering: فهرست

Care should be taken to conform to the following conventions for specific characters:

1. *Kashida*: When kashidas occur in the text, insert one kashida regardless of how many appear to be used in the original text.
2. *Long/short dash*: The long and short dash (or dash vs. hyphen) should be distinguished and matched to the original text. For handwritten documents, transcribers should use their best judgment to distinguish the two lengths, and decisions should be recorded in the documentation.
3. *Long/short underscore*: Long and short underscores should be distinguished and matched to the original text. For handwritten documents, transcribers should use their best judgment to distinguish the two lengths, and decisions should be recorded in the documentation.
4. *Parentheses*: When parentheses surround a sequence of characters or words, a single space should be inserted between the text and the parentheses, for example, (there is a single space before “there” and a single space after the final word in this sequence).
5. *Spacing*: When printed words give the illusion of having a space in between letters in the word due to typographical renditions, *do not* insert a space manually to mimic this artifact.
6. *Typing Errors*: Text should be reproduced exactly as it appears, including any errors that occur in the original text. Transcribers should not correct errors in spelling or grammar when transcribing text.

If non-Dari words or characters appear in the original text, they should be rendered exactly as they appear, and they need not be annotated to indicate the language. Character formats such as italics, bold, and underline are not required to be transcribed.

If the transcription team should come to an agreement on additional transcription conventions for a given data set, then these conventions should be coordinated prior to its implementation, and if accepted whole or in part, be made part of and documented and delivered with the transcription files.

---

4. Translation of Document Images

This section provides guidelines for the translation of text content found in document page images obtained from printed materials. Translations of the text from document image files do not preserve any features of the content layout format. Only the reading order should be preserved.

4.1 Translation Unit

The basic translation unit (TU) corresponds to a punctuated sentence. In some cases, a TU might extend to several lines of transcribed text. In other cases, phrasal units or even single words may be translation units. The following are all TUs and should correspond to separate lines in the translation file:

1. *Punctuated sentences*: Sentence punctuation is usually the period (full stop) ".", the question mark "?", the exclamation point "!", and the semicolon ";".
2. *Titles and other headings*: These are not usually full sentences, but they should be treated as TUs. Section numbers that enumerate headings should be included with the heading. Headings that span more than one line should be treated as one TU if the second line continues the first line. For example, the title and chapter in (a) are two TUs, but the two-line header in (b) is one TU.

    a. Alice in Wonderland
    Chapter 1

    b. 1.3 Representing Characters
    in the text

3. *Table cells*: Each cell in a table and each item in a list should be translated as a separate TU. The order of the cells should follow the transcription order described in section 2.2.
4. *Headers and footers*: Headers and footers may contain more than one TU. For example, if a header contains an author name flushed left and an abbreviated title flushed right, these should be treated as two TUs and ordered following the conventions described in section 3.2.

Page numbers are *not* TUs and should *not* be included in translation files.

4.2 Translation

4.2.1 Style of Translation

The goal of these translations is to take the source text and translate it, producing a result that retains the basic characteristics of the original genre. For example, news articles should be translated in news style. The translation must convey the meaning correctly and should also sound natural as it is read aloud.

The translator should try to translate as literally as possible, but without sacrificing fluency or naturalness. For example, there are many ways to translate "You were hit by a bullet." The preferred translation is one that is closest to "You were hit by a bullet" (rather than "a bullet hit you" or "it was a bullet that hit you"). If the text refers to the U.S. President as "Obama," the translation should also use "Obama" and not "Mr. Obama" or "President Obama".

If the language is idiomatic, the translator should translate the idiom's meaning rather than translating literally, e.g., the French expression *appeler un chat un chat* has an English idiomatic equivalent, *to call a spade a spade*, which should be used rather than the literal translation *to call a cat a cat*. If there are content errors in the source text, either factual or grammatical, they should be translated as is and *not* corrected in the translation. Typographical errors (e.g., misspellings) *should* be corrected if the translator is confident that it is an error in the transcription and not a content error from the source, as previously described. The corrected words should be marked with an "=" sign, for example, *Man ist, was man ist* (*You are what you are*) would be corrected to *Man ist, was man =isst* (*You are what you eat*).

The translation should maintain the same language style (or *register*) as the source. For example, if the source is polite, the translation should maintain the same level of politeness. If the source is rude or angry, the translation should be rude or angry. If the source contains a headline or title, the translation should reflect the appropriate style and capitalization for titles in the target language. Curse words and phrases should be translated with appropriately strong language.

4.2.2 Spelling Target-language Words

It is very important for translators to follow the writing conventions supplied at all times. There can be a lot of variation in how words are spelled, whether words are written as single words or broken up with one or more spaces, and so on. Ordinarily, this variation doesn't matter very much because a person reading the text will be able to figure out what is meant and will not be too confused about different spellings and spacing.

However, for our work, any variability will require a lot more effort and cost a lot more because the text will be read by computers. The computer does not understand that different spellings may mean the same word and that a "word" may or may not have a space in the middle. It is

therefore extremely important for translators to be as consistent as possible in how things are written and to proofread very carefully.

The translation team should come to an agreement on writing conventions for a given data set and come up with a word list or standard reference dictionaries for the task. These conventions should also be consistent with transcription conventions that have been established for the language and should be delivered with the translations.

4.2.3 Punctuation and Case

Punctuation should be preserved if it is synonymous in the source and the translation. If the source punctuation would be unnatural or would have a different meaning in the translation, translators should use the minimum amount of punctuation required to capture the meaning of the source.

Translators should use normal upper- and lowercase for typing English, i.e., the beginnings of sentences and proper names should be capitalized.

4.2.4 Context

Translators should always consider the context when translating. For instance, the English word "okay" could mean several things in Dari depending on the context. Translators should translate Dari words like "okay" as close to authentic English as possible.

4.2.5 Proper Names

If possible, translators should spell translated names according to conventional spellings. For example, well-known person and place names (e.g., George Bush, John, Baghdad, Kabul) should be translated as they are written normally in English.

When multiple spellings of names are possible, translators should agree to use a single spelling. If no conventional translation springs to mind, then translators should transliterate the name as best they can (i.e., write it out in the target alphabet as it is pronounced in the source language) and put the original word in brackets ("[ ]") just after the transliteration.

Some proper names may be translated partly by sound and partly by meaning. Translators should strive for the most natural translation. For example, "Southern California" may be best translated into another language by translating "Southern" but transliterating "California."

A consistent system for transliterating Dari names should be adopted, and the system should be documented and delivered with the translations. The delivery should consist of each name using Dari characters associated with the spelling of that name using Roman characters.

4.2.6 Numbers

When the source is text from a digital or paper document, numbers should be translated into the equivalent form in the target language. For example, a date in conventional English format such

as “July 2, 2010” should be translated into a similar conventional format in Dari. This also applies to other units of measurement when they are equivalent in the two languages, e.g., *kilometers, liters, or grams.* If the units are not equivalent, such as *miles, pounds,* or units of currency, an appropriate translation of the unit should be used and should not be abbreviated. For example, the English sentence “I have $5” should be translated into French as “J’ai 5 dollars.”

4.2.7 Abjad Order

If a document uses *abjad* ordering of content, the order identifiers should be translated into English letters as used in standard outline form. For example, a list using abjad order to designate alphabetical list elements is rendered as *a, b, c,* etc.

4.2.8 Errors in the Original Document

Although transcription should reproduce the original text exactly as it appears, translations should not preserve errors. Spelling errors in the original document should be corrected in the Dari text of the translation document, but errors that would require more complex re-writing should not be corrected. However, sentences with grammatical errors should be translated as if they do not contain any errors.

4.3 Translation Format

Each document image file should correspond to a transcription file with the same name, except for the file format suffix, which should be “_trl.txt”. For example, if the document image file is named “file.pdf”, the transcription file should be named “file_trl.txt”.

Lines in the file should have the following format where [tab] represents a 2-space tab delimiter and square brackets represent positions in the line (not included in the transcription). The translation unit is described in section 4.1.

[image filename][**tab**]Line 0001[**tab**]first TU of file from transcription
[image filename][**tab**]Line 0001[**tab**]translation of first TU from transcription
[image filename][**tab**]Line 0002[**tab**]second TU of file from transcription
[image filename][**tab**]Line 0002[**tab**]translation of second TU from transcription

and so on until all of the transcribed TUs are translated.

The image filename should match the filename of the document image. If the TU begins in one document image and ends in another document image, the image filename that is used should be the name of the file in which the TU begins. If the TU begins at the end of a document image and there are no subsequent images from the same original document, the partial TU should be translated as a partial translation.

Translation of tables should preserve the order of the transcription, but each cell should be on a separate line that ends with a carriage-return/linefeed. Example 8 illustrates the transcription and translation of a table.

Example 8: Translation for Transcription of Table from Example 6

TRANSCRIPTION

Dari Table Title on Top
Dari Table Column Label 3**[tab]**Dari Table Column Label 2**[tab]**Dari Table Column Label 1
Dari table cell 3**[tab]**Dari table cell2**[tab]**Dari table cell 1
Dari table cell 6**[tab]**Dari table cell 5**[tab]**Dari table cell 4
Dari Table Caption Below

TRANSLATION

Dari Table Title on Top
Translation of Dari Table Title on Top
Dari Table Column Label 1
Translation of Dari Table Column Label 1
Dari Table Column Label 2
Translation of Dari Table Column Label 2
Dari Table Column Label 3
Translation of Dari Table Column Label 3
Dari table cell 1
Translation of Dari table cell 1
Dari table cell 2
Translation of Dari table cell 2
Dari table cell 3
Translation of Dari table cell 3
Dari table cell 4
Translation of Dari table cell 4
Dari table cell 5
Translation of Dari table cell 5
Dari table cell 6
Translation of Dari table cell 6
Dari Table Caption Below
Translation of Dari Table Caption Below

5. Conclusions and Recommendations

This report presents a set of guidelines for transcription and translation of foreign text regions found in printed material specific to the Dari language. Although limited here only to address the document layout text content only, it provides an annotation schema for facilitating the transcription and the generation of respective translation of text found in document images as well as a discussion on language-specific features and issues. It reflects our effort at adopting a

common set of annotations usable for generating transcription and equivalent translations text for use in OCR and MT processes by facilitating an approach conducive to easy parsing and sharing in extensible markup language (XML)-based systems for use in technology developments.

As the information age moves forward to enable automation, desktop publishing and printing will continue to spread throughout the world. As the need to process more printed materials in foreign languages and associated scripts grows, the requirement for representative data set generation will continue to exist to support research, development testing, and experimentation activities. Guidelines to develop such resources offers a systematic approach to capture and exchange necessary information needed to satisfy a specific goal or objective.

6. References

1. Hernandez, L.; Schlesiger, C. OCR for Collection, Management, and Retrieval of Documents: Development and Trial of a Documentation Exploitation Suite. *Proceedings of 2003 Symposium on Document Image Understanding Technology*, University of Maryland, Laboratory for Language and Media Processing, 2003.

2. Zavorin, I; Borovikov, E; Borovikov, A; Hernandez, L, Summers, K; Turner, M. A Multi-evidence, Multi-engine OCR System. *In Proceedings of SPIE/IS&T Electronic Imaging conference on Document Recognition & Retrieval XIV, 2007.*

3. Morgan, J, J. *Human in the Loop Machine Translation of Medical Terminology*, ARL-MR-0743; U.S. Army Research Laboratory; Adelphi, MD, 2010.

4. Giffen, N.; Hernandez, L.; Briesch, D. "Experimental Laboratory Environments: Image and OCR Tool Kit (IOTK) Utility Exploration," in *ARL Summer Student Research Symposium Volume 1: Compendium of Papers*; ARL TM-2008; U.S. Army Research Laboratory: Adelphi, MD, 2008, 5–53.

List of Symbols, Abbreviations, and Acronyms

MT	machine translation
OCR	optical character recognition
pdf	portable document format
TRS	transcription
TU	translation unit
txt	text file extension
UTF-8	Unicode Transformation Format 8
XML	extensible markup language

No.
of Copies Organization

1 (PDF only) DEFENSE TECHNICAL INFORMATION CTR
DTIC OCA
8725 JOHN J KINGMAN RD
STE 0944
FORT BELVOIR VA 22060-6218

1 DIRECTOR
US ARMY RESEARCH LAB
IMAL HRA
2800 POWDER MILL RD
ADELPHI MD 20783-1197

1 DIRECTOR
US ARMY RESEARCH LAB
RDRL CIO LL
2800 POWDER MILL RD
ADELPHI MD 20783-1197

1 DIRECTOR
US ARMY RESEARCH LAB
RDRL CIO LT
2800 POWDER MILL RD
ADELPHI MD 20783-1197

5 HCS DIRECTOR
US ARMY RESEARCH LAB
RDRL CII
B BROOME
RDRL CII T
V M HOLLAND
L HERNANDEZ
D BRIESCH
S LAROCCA
2800 POWDER MILL ROAD
ADELPHI, MD 20783-1197

2 HCS THE MITRE CORPORATION
SHERRI CONDON
DAN LOEHR
7525 COLSHIRE DR MAILSTOP H305
MCLEAN, VA 22102